A PROTOCOL TO FIX BROKEN LINKS ON THE WORLD WIDE WEB

BACKGROUND OF THE INVENTION

Field of the Invention

The invention relates to client-server data communication systems. More particularly, it relates to repairing links between pages in a client-server network.

Description of the Related Art

The Internet, as it is popularly known, has become an important and useful tool for accessing a wide variety of information. One component of the Internet is the World Wide Web (hereinafter the web). In recent years the web has become an increasing popular vehicle for providing information to virtually anyone with access to the Internet. Many websites have been established to provide, over the web, information in many different forms, such as text, graphics, video and audio information.

A typical website is hosted on a network server computer that includes application software programs. The server, also known as a web server, is connected to the Internet. By connecting the web server to the Internet, clients that are connected to the Internet can access the website via the web server. Usually, a client is located remotely from the web server, although the client and the server can be at the same location. A web server also can be connected to a private intranet, as opposed to or in addition to the public Internet, in order to make a website privately available to clients within an organization.

A client-server communication system used on the web is shown on Fig. 1. The system includes a client 2 that is connected to a monitor 4 and to a network 6, such as the Internet. The client sends and receives messages over the network 6 to a server, such as web server 8, web server 10, or web server 12 shown in Fig. 1. The web servers host web sites that include one or more web pages. For example, web server 8 hosts web site 13 that contains a series of web pages 14a through 14n. Similarly, web server 10 hosts web site 15 that contains a series of pages 16a through 16n, and web server 12 hosts web site 17 that contains a series of pages 18a through 18n. A uniform resource locator (URL) is a string that gives information about the location of a particular resource (such as a file, image, or program) on the Internet. Generally, each web page has a unique URL.

The client 2 accesses web pages on a website by using a web browser 20; that is, a software program that runs on the client and receives from the server information formatted in a known manner. A very popular format for information sent over the web from a server to a client is the Hyper Text Mark-up Language (HTML).

A web server typically takes user input, in the form of a URL, and returns the file(s) that correspond to that web page. This process begins by the client browser sending a request to a web server indicated by the URL. Once the web server receives the client request, it locates the file, or executes the program, specified by the URL and sends the file back to the client browser. The file(s) making up the web page that has been delivered to the client is held in a cache memory for use by the browser 20. Web page 22 shown in Fig. 1 represents a web page that is stored in the browser's cache. The browser interprets the HTML code in the web page to generate a display 24 on monitor 4. If the web server encounters a problem while processing the client browser's request, it returns an error code.

One web page on the Internet can reference another web page on the Internet through the use of URL links. These links are basically URL strings contained within special HTML tags. When the user clicks on such a link the client browser requests from a web server the resource specified by the URL and displays that resource, such as an HTML web page file, on the client browser. Here, for purposes of illustration, referring to Fig. 1, assume the web page 22 held in the browser's cache and displayed as page 24 came from web site 13 in web server 8. The web page display 24 includes two hypertext links to other web pages held on different severs. URL link B ("link B") 26 contains the URL of web page 16a stored in web server 10. URL link C ("link C") 28 contains the URL of web page 18a stored in web server 12. If a user selects link B 26, the browser sends a message to the web server 10 to return the web page corresponding to the URL of link B. Here, web server 10 returns an HTML copy of web page 16a to client 2. Similarly, if a user selects link C 28, the browser sends a message to the web server 12 to return the web page corresponding to the URL of link C. Here, web server 12 returns an HTML copy of web page 18a to client 2.

A commonly encountered problem with many web pages is that the hypertext links on those pages might become stale, or broken, such that the URL within the hypertext link no longer refers to the location of a web page. The problem of broken links, also known as linkrot, occurs commonly on sites and pages throughout the web. Web surfers find broken links to be annoying and usually tend to avoid sites that have many broken links. For web page authors, fixing broken links can be tedious and labor intensive.

A URL link can be considered to be broken when, for example

1. The file specified by the URL has been renamed in the web server.

2. The file specified by the URL has been deleted in the web server.

3. The location of the file under the web browser is changed.

Under any one of these circumstances the web server returns an error message (e.g., error code 404) back to the client browser.

Broken links are very annoying to the users and are quite common on the World Wide Web. A 1997 World Wide Web user survey rated broken links to be the most frequent problem encountered by users.

Fixing broken links is a significant inconvenience for web developers. It is a task that is carried out manually, and hence, is labor intensive and time consuming. Despite the fact that broken links are regarded as one of the most serious problems on the World Wide web, no definitive solutions that solve the problems once and for all has yet been developed.

Proposed solutions to date are difficult to implement and do not operate automatically. One such solution recommends web developers follow rules, set forth below, to prevent broken links.

1. Check the web page links frequently and fix them to reduce outbound linkrot.

2. Keep old pages on the server forever and if moving pages place a redirect link on the old page.

Web developers often either are not aware of such rules or simply do not follow them, as illustrated by the large number of broken links on the World Wide Web. Accordingly, there is a long felt but as of yet unsolved need to automatically detect and fix broken links.

SUMMARY OF THE INVENTION

Therefore, in light of the above, and for other reasons that will become apparent when the invention is fully described, an object of the invention is to automatically fix broken links on web pages.

Another object of the invention is to automatically detect broken hypertext links.

A further object of the invention is to correct hypertext links in documents without requiring modifications to client software.

Yet another object of the invention is to enable a service for automatically detecting and correcting hypertext links embedded in documents on a web site.

A still further object of the invention is to provide a protocol for detecting, and correcting or removing broken hypertext links.

The aforesaid objects are achieved individually and in combination, and it is not intended that the invention be construed as requiring two or more of the objects to be combined unless expressly required by the claims attached hereto.

In accordance with the invention, a protocol is described here that can be used to fix broken links automatically, thereby saving countless hours wasted by web surfers trying to navigate using a broken link or web authors trying to fix broken links to their web pages.

The above and still further objects, features and advantages of the invention will become apparent upon consideration of the following descriptions and descriptive figures of specific embodiments thereof. While these descriptions go into specific details of the invention, it should be understood that variations may and do exist and would be apparent to those skilled in the art based on the descriptions herein.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram of a client-server network system.

Fig. 2 shows a client-server network system with link checking and link correction services.

Fig. 3 illustrates a link mapping table.

Figs. 4A and 4B are flow diagrams for describing link correction and link checking services.

Figs. 5A and 5B illustrate hyperlink validation protocol (HLVP) request and response messages.

DETAILED DESCRIPTION

The embodiments described below are described with reference to the above drawings, in which like reference numerals designate like components.

A solution to the problems associated with fixing broken web links can be based in a web server that operates automatically to correct or remove broken links. One aspect of the solution checks hypertext links in a web page to determine whether or not those links are broken. If they are broken then the web page is automatically corrected so that the link refers to the correct URL, or the link is removed. As described in more detail below, checking for broken links and fixing them can be accomplished by including the following extensions to a web server, without the need to modify any clients, although using such extensions in a client is not precluded.

1. A mapping table that maps a new URL to the old URL.

2. A hyperlink validation protocol (HLVP).

3. A link correction service.

A local server, such as a web host that hosts one or more web sites, can include a link correction service that performs link correction on web pages, or other types of documents that have broken URLs in those sites. The link correction service operates to correct broken hypertext links in the document and produce a corrected, or fixed document to use with the web site. The link correction service employs a new protocol that is described in more detail below for validating hyperlinks in a document. Based on a URL in the hypertext link, the link correction service identifies a remote web server that is referred to by the URL in the document. The link correction service uses the link verification protocol to request the remote web server to validate the link. The link correction service can reside in the web server that is attempting to the fix broken links in the documents that it manages, or alternatively the service can reside in a client.

A client-server network system that performs link checking and link correction services is illustrated in Fig. 2. Referring to Fig. 2, a web server 8 includes a link correction service unit 30 that performs link correction on web pages or other types of documents. The box labeled 32 in Fig. 2 represents one or more documents that contain hyperlinks or URLs pointing to various resources that may be present on one or more remote web servers. For purposes of illustration here, it is assumed that many of the hyperlinks in document 32 are broken. The link correction service unit 30 operates to read and parse the document 32, and then fix or remove any of the broken hyperlinks found in that document. The link correction service unit 30 upon fixing or removing the broken hyperlinks, generates a modified form of document 32 where the modifications can include: 1) either broken links being replaced by

fixed links, or 2) broken links being removed from the document. The document modified to correct or remove broken links is shown as the box labeled 34 in Fig. 2.

A link checking service informs the link correction service whether a hyperlink that points to the web server where the link checking service resides is: 1) valid, 2) no longer exists, 3) has been replaced by an alternate URL, or 4) has an unknown status. The link checking service typically operates on another web server that, generally, is remote from the web server running the link correction service, although they can reside on the same server. A remote web server, such as web server 10 shown in Fig. 2, includes a link checking service unit 36 that checks the validity of a link based on a request from the link correction service unit 30. Within the remote web server are a set of documents that are managed by web server 10. Often such documents are in HTML form, and accordingly, box 38 in Fig. 2 is labeled HTML file repository, although documents in other formats can reside in repository 38. The remote web server 10 also includes a mapping table 40 for mapping old URLs to new URLs, and a mapping tale maintenance unit 42 for maintaining the mapping table.

The mapping table 40 lists the changes that occur in the URL of the resources that the web server exposes to the outside world. A change in the URL of a resource has a corresponding entry in the mapping table. An example of such a mapping table is shown in Fig. 3. Here, the mapping table 52 includes the old URLs, listed in column 54; codes indicating the status of the URL, listed in column 56; and the new URLs corresponding to the changed locations, listed in column 58. The mapping table can be stored either in a file, in a database, or in any other suitable form that allows access to and use of the mapping table. The table can be generated and updated manually by a webmaster or through the use

of a software utility program running in the mapping table maintenance unit 42. The mapping table maintenance unit 42 can operate with other software programs that determine an alternate URL for a given URL, to automatically update the mapping table.

In the mapping table 52, shown in Fig. 3, five rows 59a-59e are shown corresponding to five files that are stored, or were previously stored, in remote web server 10. For example, row 59a of the mapping table indicates that the file having a URL http://www.company-name.com/removed.html has been removed since it has a URL status code of "-1". Row 59b indicates that the file with URL http://www.company-name.com/index.html still resides where the URL indicates it can be located, and accordingly, has a URL status code of "0". Row 59c indicates that the file with URL http://www.company-name.com/moved.html has moved to a new location since its URL status code is "1". Row 59c also indicates the new URL for the file, namely, the URL http://www-3.company-name.com/moved.html.

Other remote web servers, such as web server 12 shown in Fig. 2, can include a link checking service. Those other web servers, such as web server 12, include a link checking service unit 44, an HTML file repository 46, a mapping table 48 and a mapping table maintenance unit 50.

The link correction and checking services preferably are implemented in software with program instructions being executed on a computer to operate as described below with reference to Figs. 4A and 4B. Referring to Fig. 4A, the link correction service unit operates by reading a document that may possibly contain one or more broken links 60. The link correction service unit checks each of the hyperlinks found in the document one by one. Upon inputting the document, the link correction service parses that document to obtain the

first hyperlink or URL to be checked 62. For each hyperlink or URL in document A, the link correction service unit accesses the web server indicated by the URL and sends a request to the link checking service unit in that server to determine if the link is valid 64. Operations performed by the link checking service are illustrated in Fig. 4A within box 66.

In response to the link checking service unit at the remote server receiving the request for validating a link, it checks if the link is valid by determining if the document pointed to by the URL is actually present among the documents managed by its web server 68. If the document is found to be present, then the link checking service returns the code "0" to the link validation service unit, indicating that the link is valid. The link correction service unit then obtains the next link in the document to be checked 70, and then checks that next link in a similar manner. If the document is not found among the documents managed by the web server, the link checking service unit uses the mapping table to determine if an alternate link is available 72. If no alternate link is available a code "-1" is returned and if the link cannot be found in the mapping table a code "-2" is returned. Upon receiving either of those codes the link correction service unit removes the link from the document 74 and obtains the next link in the document to check 70. If an alternate link is available, the link checking service unit returns a code "1" and the URL of the alternate link. The link correction service unit modifies the document to replace the broken link with the alternate link 76 and then obtains the next link in the document to check 70.

Fig. 4B illustrates a process performed by the link checking service in the remote web server. The process begins by the remote web server receiving a URL to check from a link correction service unit 78. The link checking service unit parses the URL to determine the local path within the remote web server 80, and determines if a file corresponding to that

URL exists in the remote web server 82. The link checking service unit can consult the mapping table to determine if the requested URL is listed in the mapping table with an indication that the URL is valid. Alternatively, the link checking service unit can determine through the remote web server's file system if a path to the requested file exists, and hence, the URL is valid. If the URL is valid, a return code "0" is returned to the link correction service 84.

If the URL does not correspond to a file that is found among the documents managed by the remote web server, the link checking service consults the mapping table to determine if an alternate URL is available for the hyperlink. A mapping table can show the status of a URL to be valid (status code "0"), or as being permanently removed (status code "-1"), or it can show a new or alternate URL for the URL that the link checking service is attempting to validate (status code "1" + alternate URL). The link checking service unit determines if the URL corresponds to a file that has been deleted 86. If the mapping table indicates that the file corresponding to the URL has been deleted, a return code of "-1" is returned 88.

If the file is not indicated in the mapping table as being deleted, then the link checking service determines if the mapping table indicates that the file corresponding to the requested URL has been moved 90. If so, a return code of "1" and the alternate URL indicated in the mapping table are returned 92.

If the document corresponding to the URL is not found among the documents being managed by the web server and if there is no entry in the mapping table also for that URL, then that URL is considered to be invalid. This condition can occur as a result of a link correction service sending a request to a link checking service to validate a false hyperlink,

or a hyperlink believed to be incorrect. In such a situation the link checking service responds by returning a status code of "-2" 94.

For each link in a document, the link correction service might have to send requests to various link checking services running on different remote web servers, since hyperlinks contained within a document can point to different web servers.

For the purpose of description below assume the following parameters set forth in Table 1.

Table 1

Domain name of web server:	www.company-name.com
Root directory of web server:	/
Requested file name:	foo.html

Web servers today implement the Hypertext Transfer Protocol (http) for communicating requests and responses for web pages between a client and a server. In order to validate a link, the web servers 8, 10 and 12 shown in Fig. 2 are configured to operate according to a new protocol, referred to here as a HyperLink Validation Protocol (HLVP). The HLVP protocol has the following characteristics.

1. A protocol request is URL based, similar to HTTP. To validate the http link http://www.company-name.com/foo.html a client program sends an HLVP request message 96 in the form shown in Fig. 5A, where "hlvp" indicates that the request follows the HLVP protocol, and the URL to be validated is "www.company-name.com/foo.com". Note that

the URL for the requested document is exactly the same as in HTTP except for the name of the protocol in the beginning of the URL string ("hlvp" as opposed to "http").

2. An HLVP protocol response message 98 validates the URL and has the form shown in Fig. 5B. The HLVP response message includes two fields, a URL status field 98a and a URL field 98b. The URL status field includes one of the numeric codes shown in Fig. 5B depending on the determined status of the URL. For example, if the URL is determined to be valid, the URL status field of the HLVP response message contains the response code "0". If the location of the URL is determined to be changed, the URL status field of the HLVP response message contains the response code "1" and the URL field contains the new URL identifying the present location of the document.

The link correction service first checks all the web pages stored on the server to determine all the external links in those pages. Then, the link correction service uses the HLVP protocol to validate the links on the web pages. The actions initiated by the link correction service based on the HLVP response is summarized as follows.

1. If an HLVP response for a link on the web page has a URL status code of "0" then that link is valid and the link correction service takes no further action.

2. If an HLVP response for a link on a web page has a URL status code of "-1" then the web page pointed to by the link no longer exists. In that case, the link correction service edits the web page to remove the broken link.

3. If an HLVP response for a link has a URL status code of "1" indicating a change in the location of the file corresponding to the URL, the link correction service edits

the web page to change the broken URL to point to the new URL contained in the HLVP response message.

4. If an HLVP response for a link on a web page has a URL status code of "-2" then the link checking service does not know about the requested URL. Accordingly, the resource pointed to by the link is unknown by the remote web server. In that case, the link correction service edits the web page to remove the unknown link.

The systems and techniques described here for fixing broken links do not require modifying client web browsers, but rather are based in the web server. This solution when implemented in a web server configured as described above, allows the web server to autonomously check, validate, and correct the URLs present in the web pages that the web server manages.

Having described embodiments of apparatuses, articles of manufacture and methods of correcting broken hyperlinks, it is believed that other modifications, variations and changes will be suggested to those skilled in the art in view of the teachings set forth herein. It is therefore to be understood that all such variations, modifications and changes are believed to fall within the scope of the present invention as defined by the appended claims. Although specific terms are employed herein, they are used in their ordinary and accustomed manner only, unless expressly defined differently herein, and not for purposes of limitation.